edWisor

Advanced Predictive Analytics Using R

- MUQUAYYAR AHMED
DATA SCIENTIST

We learnt!!!

Statistics

Predictive Analytics

Advanced Predictive Analytics

- Introduction to Machine learning

Today's Agenda

- Decision Tree

What is Decision Trees??

- A predictive model based on a branching series of Boolean tests
- Can be used for classification and regression
- There are number of different types of decision trees that can be used in Machine learning algorithms

Contd..

- Decision tree is a rule. Each branch connects nodes with “and” and multiple branches are connected by “or”.

- Extremely easy to understand by the business users.

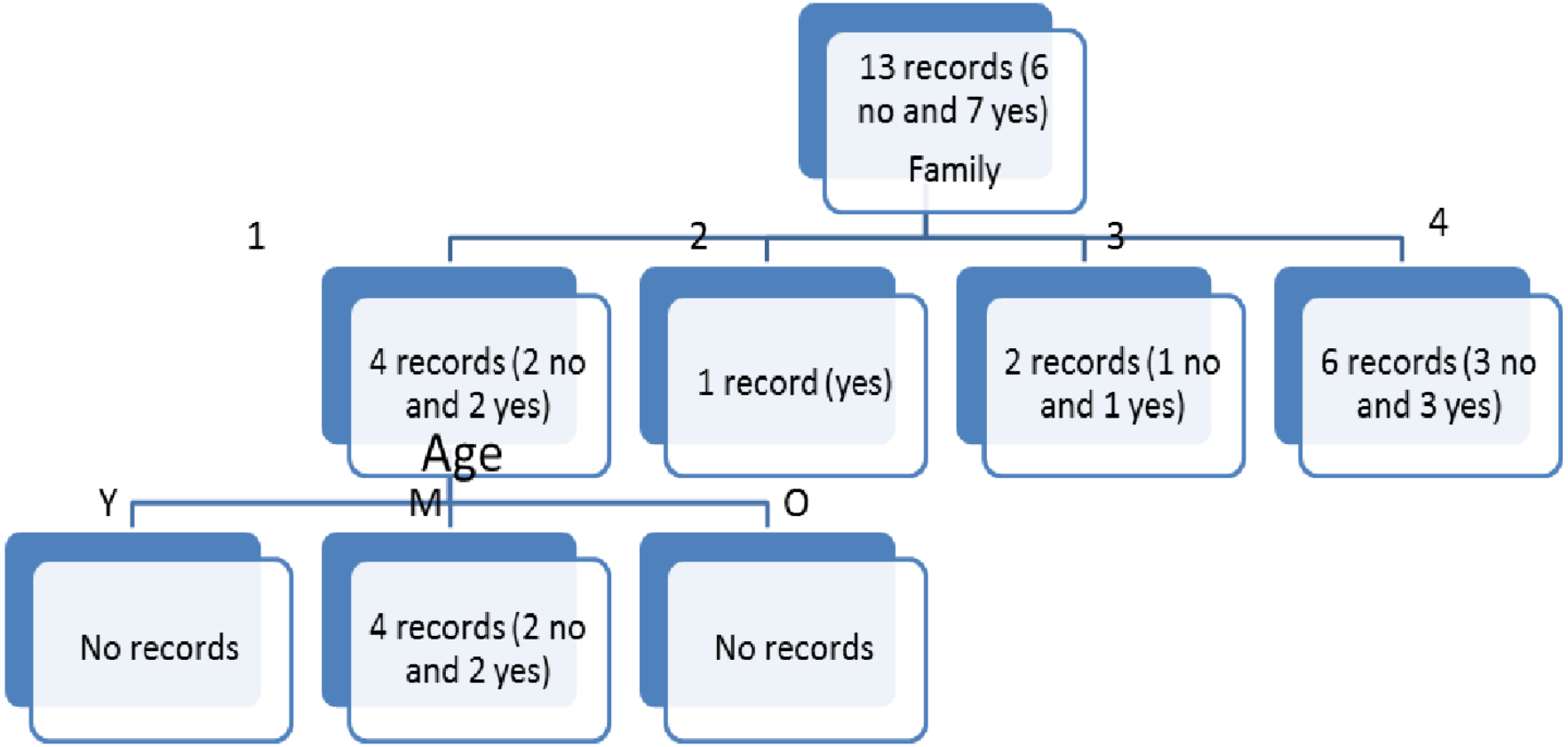
- Build some intuitions about your customer base. E.g. “are customers with different family sizes truly different?”

edWisor

Sample Data

ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1

Decision trees with different attributes



Sample Rules

- If CCAvg is medium then loan = accept
- If CCAvg is low and income is low, then loan is “not accept”
- If CCAvg is low and income is high, then loan is accept
- If (CCAvg is medium) or(CCAvg is high) or(CCAvg is low and income is high) then loan = accept

Two most popular decision tree algorithms

- C5.0
  - Multi split
  - Information gain
  - Rule based pruning

- CART
  - Binary split
  - Gini index
  - Tree based pruning

Information Gain

- It represents the expected amount of information that would be needed
- Measure of purity
- Loss of entropy
- IG = Entropy of the system before split –

  Entropy of the system after split

- Entropy: Uncertainty in the data/Measure of impurity, can be calculated as

$$H = -\sum_i p_i \log_2 p_i$$

- Selects the variable whose Information gain is high

Let us understand data..

•Let us consider a data set of 70 people (30 men and 40 women). Let us assume that their ages are spread from 1-99. 30 men have ages ranging from 1 to 40. The age of the 40 women range from 41 to 99.

Men

Number	Age	Sensible
Man 1	1	No
Man 2	2	No
...	...	...
Man 40	40	No

Female

Number	Age	Sensible
Woman 1	41	Yes
Woman 2	42	Yes
...	...	...
Woman 59	99	Yes
Woman 60	99	No

Contd..

- Our training set consists of last 20 males (all insensible and age between 21-40) and last 30 females (all sensible and age 71-99 except the last one). Testing contains the other half where all women are sensible.

Hypothesis

If it is a man: No sensible decision

If it is a woman: Sensible decision (the only loan case where she does not must be noise.

Information Gain!!

edWisor

Entropy before splitting the variables

Number of sensible people: 29
Number of insensible people: 21 (twenty males and one female)
Entropy of the system:

$\sum -p_i \log_2 p_i$=-(29/50*LOG(29/50,2))-(21/50*LOG(21/50,2))= 0.9814

Entropy after split

Split based on **gender**

- Now, the system will have two sub systems (male and female).
- Entropy of the male system = (20/20*LOG(20/20,2) = 0 Entropy of the female system = 29/30*LOG(29/30,2) + 1/30*LOG(1/30,2) = 0.2108
- Entropy of the total system after split is the weighted average of the individual parts

  = 20/50 (Male system) + 30/50 (Female system)

  = 3/5*02108 = 0.1265

Entropy loss or information gain= 0.9814-0.1265 = 0.8549

If Rules then which Rule??

- Support
  - How frequently the item-set appears in the database
- Confidence
  - Confidence is an indication of how often the rule has been found to be true.
- Lift
  - Ratio of the observed support to that expected if X and Y were independent